Privacy Rules: Approach in the Label or Textual Format

Sergio Donizetti Zorzo
Computer Science Department
Federal University of São Carlos
zorzo@dc.ufscar.br

Diego Henrique Dias
Computer Science Department
Federal University of São Carlos
diegohdias@hotmail.com

Diego Roberto Gonçalves de Pontes
Computer Science Department
Federal University of São Carlos
diego.pontes@dc.ufscar.br

José Santiago Moreira de Mello
Computer Science Department
Federal University of São Carlos
jose.mello@dc.ufscar.br

Abstract

Users usually don't read privacy policies of the websites accessed. This paper presents the privacy policy of the websites in a format named Privacy Label for being similar to nutritional labels. It is presented on the standardized-table format of items of privacy policies, including governmental policies. This format was compared to the policies described as full text written in natural language based on the perception of 198 participant students of the different areas. The results indicate that the Privacy Label format facilitates users' comprehension of the policy content and made them more aware of elements that they would usually dismiss when reading a textual privacy policy.

Keywords

privacy, data privacy, human-computer interaction, IS policy, policy evaluation.

Introduction

Data from web users are very important to companies that offer virtual services and are commonly used to promote specific ads, get market statistics, shopping habits and products recommendations. All this information may be collected upon signing up to a service, the use of cookies, the click tracking and the pages visited by the user.

When an online service collects users' data, the companies need to provide information regarding which data will be collected and what will do with it, according to what it is demanded by the regulatory agencies (as OECD, FTC, et al). A company's privacy policy is considered appropriate if their target audience is able to understand it easily Jensen and Potts (2004), and that strongly depends on the reading and comprehension abilities of the final user. Though regulatory agencies define writing patterns to be used in privacy policies, some companies don't follow such patterns. By doing that, they can make their policies long and hard to understand for some users.

According to Jensen and Potts (2004), most users do not give enough attention to privacy policies either for not knowing technical terms or the long time spent on reading it. Usually, they simply accept the policies, not knowing what kind of data is being collected. The ideal situation is that the user must read, understand and then decide to accept or not the privacy policy of a given online service.

Kelley et al. (2009) observed the many ways that privacy policies are presented to users, and based on that, developed a graphical pattern for privacy policies, which they called privacy labels. The privacy label was inspired on nutrition labels, which help a customer decide if that product is good for him. According

to the authors, if nutrition labels can influence a customer's decision of buying a product, then privacy labels can help by giving the user a more comprehensive and easier way to understand a privacy policy than by just reading the terms.

This approach extends the work proposed by Kelley et al. (2009), adding a new field (OPT-LAW), which indicates situations requiring collection and storage of data by accounts of government laws or regulatory agencies. These legal requirements may differ depending on the country of origin. In Brazil, the law number 12,965/14 (Civil Internet Marco) defines rules for companies' use and collection of user's data, including handing such data to government entities when necessary. An additional contribution in this work is the inclusion of the icon "**Ñ**" in the table indicating that the item of privacy policy is not informed. This way to present the privacy policy was compared with the traditional one by the users' understanding of the written privacy policies and privacy labels, assessing their perception regarding data collection. The users chosen in this experiment were students of different fields from two college institutions.

In the next section, we present the works related to the proposal in this paper. To validate our privacy label proposal, which is described in the section named Privacy in the Label Format, we collected privacy policies from two companies that provide online services. These services were the most accessed by the students in the research domain, and they are related in the Methodology section. The Results Analysis section presents the results and discusses them and in the Conclusion section, we give our final remarks.

Related Works

The works related in this paper are about studies regarding users' understanding of privacy policies, and the way that they are shown to users. The work of Jensen and Potts (2004), McDonald et al. (2009), and Kelley et al. (2009) (2010) have already addressed this issue, but this proposal has the research question on how the Privacy Label format can facilitate the users' comprehension in the context of Brazil. This proposal was sponsored by the Internet Civil Marco from Brazil, and it was necessary to expand previous works.

Jensen and Potts (2004) assessed the users' understanding of technical and legal terms presented on privacy policies from online services. They observed that even though users read and agreed to the privacy policies terms, they did not have enough knowledge to be fully aware of the consequences of accepting such terms.

McDonald et al. (2009) assessed three different types of privacy policies presentations, which were natural language written privacy policy, automatically generated privacy policy and warning layers. The results showed three important observations. First, the warning layers approach suggested that the users were restricted to the complete privacy policy when they could not find the information they were looking for on the layers; besides that, warning layers may hide relevant information and reduce transparency. Second, the participants had trouble to extract concepts regarding some terms used in the policies. And third, the writing standard formats that are still under development may cause ambiguity, which can cause companies with identical services seem different for the final user.

Kelley et al. (2009) developed a table based on nutritional food charts that uses a matrix representation to identify standard items related to privacy, using a pre-defined score to rank them. The first row contains the kinds of services that may collect data. In the center on the table, regions are marked in different colors representing what kind of data is being collected and if the user can or cannot do something about it. To generate privacy policies in this table format, the authors used files from the Platform for Privacy Preferences[1] (P3P).

Using privacy tables, Kelley et al. (2010) presented a comparative study of privacy policies in table format and textual format for some online services. They showed that the understanding of privacy policies was more efficient when the information was presented in a table-wise manner for the assessed services. However, they noted that the use of a table is not the sole responsible for users to better grasp what kind of data is being collected; rather, a privacy policy written in a simple and comprehensive way, with glossary and lack of technical jargons can help users understand it in a much better way.

[1] http://www.w3.org/P3P/details.html

In this paper, changes were made in the table proposed by Kelley et al. (2009) and are presented in the section Privacy in the Label Format. This new approach is similar to the work of Kelley et al. (2010), focusing on individual questions of convergence and divergence of the understanding of the policy expected by the final users.

Privacy in the Label Format

The current way of presenting privacy policies is considered, according to Howard Beales[2], a poor mechanism of notification, with long and exhaustive texts, full of technical jargons, which are usually hard for users who decide to read the policy to understand.

The study conducted by Kelley et al. (2009) proposed a different approach to represent a privacy policy, named privacy label. This label aimed to represent in a simple and intuitive manner which data were collected from users and why.

In our study, we assessed the perception level of users, comparing the privacy label approach with the traditional way of presenting a privacy policy. To conduct our experiment, two companies were selected, following the methodology proposed in the section Methodology. Privacy policies in the label format generated from the written policies of these companies are shown in Figure 1 and Figure 2.

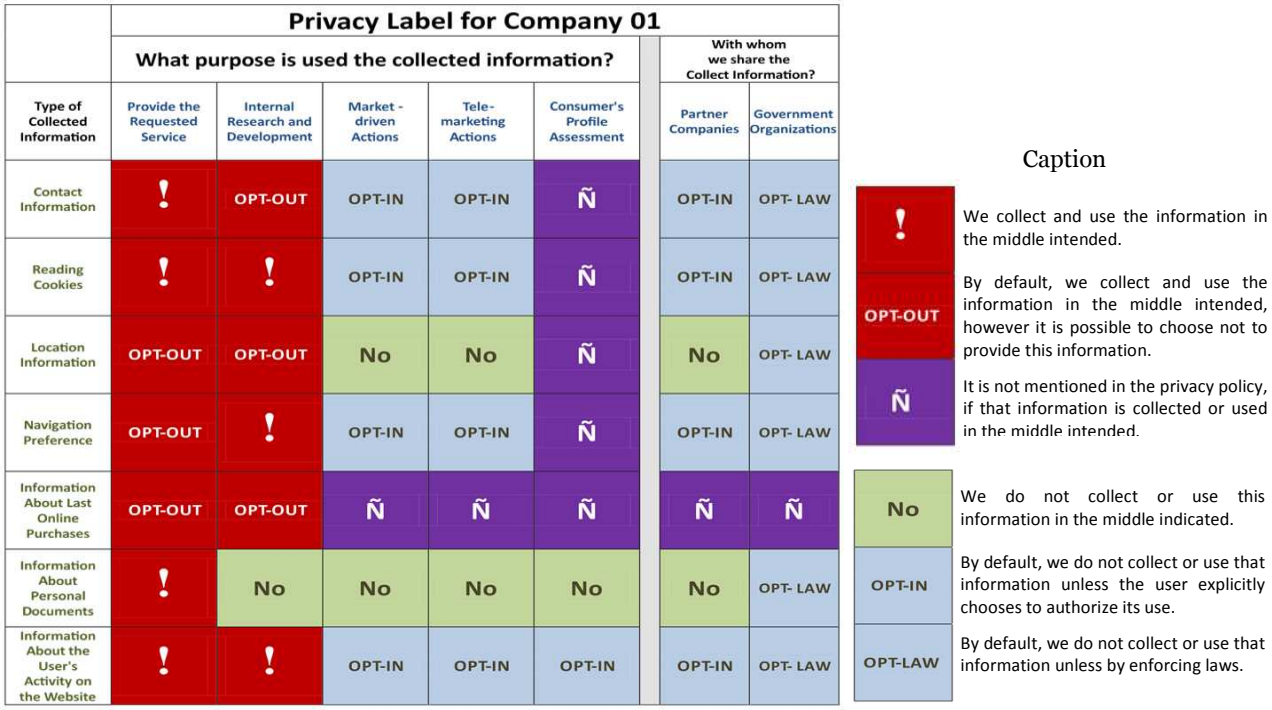

Figure 1: Privacy Label for Company 01

[2] http://www.ftc.gov/public-statements/2002/01/privacy-notices-and-federal-trade-commissions-2002-privacy-agenda

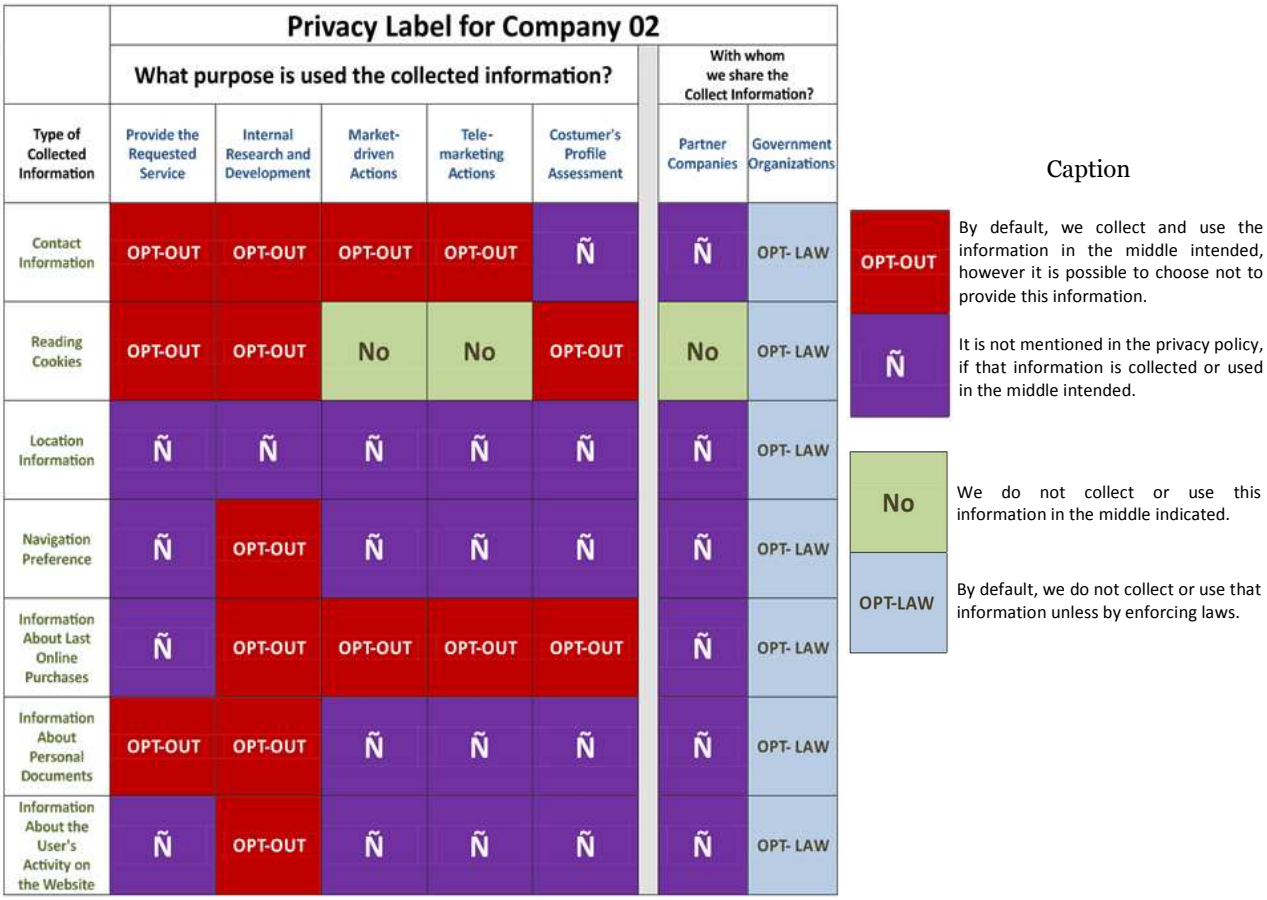

Figure 2: Privacy Label for Company 02

The privacy labels were defined in a manual process following the methodology proposed by Kelley et al. (2009). Areas colored in red indicate situations where user data will be collected. In some of these situations, the user may intervene and opt not to share his data (OPT-OUT). In other situations, this OPT-OUT is impossible since there are no available mechanisms made by the companies to allow user control over the data collection. Areas colored in blue indicate situations where either data will be collected only if the user wishes to do so (OPT-IN) or data will be collected by legal reasons (OPT-LAW) to comply with government laws or regulatory agencies, which makes the companies obliged to collect and keep user data.

In Brazil, where this research was conducted, law number 12.965/14 (Internet Civil Marco)[3] defines rules for companies that use and collect user data, including handing such data to government entities when necessary.

Areas colored in green indicate situations where the privacy policy explicitly declares that there is no data collection in a given situation. Finally, areas colored in purple indicate situations where the privacy policy does not declare if there is data collection. The lack of this information alienates users as to if the company is or is not collecting a given information categorized in the privacy label.

The purposes of each data being collected, presented in the privacy labels, were grouped in five main categories. Such categories were selected based on the frequency they were cited in the companies' written privacy policies.

The selected categories were:

3 Full document available on: http://www.camara.gov.br/sileg/integras/912989.pdf

- Provide the requested service;
- Internal Research and Development;
- Market-driven actions;
- Telemarketing actions;
- Customer's profile assessment.

Beyond these main categories, two special ones related to sharing data with third-parties were added to the label. The criterion for choosing these special categories was the same used for the main ones. The special cases are:

- partner companies
- government organizations

the different types of collected data also went through a grouping process to simplify and standardize the terms used in written privacy policies.

The methodology used in this grouping was the same proposed by Kelley et al. (2009) where information with similar content was combined under a name who could describe its function in a generic way. This information was grouped based on:

- Combined contact data such as phone number and e-mail address in one simple category named contact information.
- Combined geographical data such as city, address and place of birth in one simple category named localization information.
- The navigation preferences category resulted from grouping information such as web history, search history and other data related to the user's web surfing.
- Combined data on the place of purchase and visualized and bought products within a category named information about last online purchases.
- Combined data such as browser version, IP address, click tracking and login ID in the analyzed websites as part of a category named information about the user's activity on the website.

Methodology

This section describes how the experiment with the labeled and written privacy policies, and the questionnaire were conducted. In the first one, the selection of websites to collect the access registries was made. After this, we chose students from different undergraduate courses, where the selection method was the availability of each group to participate in the experiment. The prepared questionnaire to be used to this work was based on the guidelines proposed by Lazar et al. (2010). In the sequence, we apply the experiment using the directives presented by Lazar et al. (2010). Finally, the results were compiled; we made the analysis, and some conclusions were made.

Site Selection

The selection of the websites to collect the access registries was composed by: (1) Choosing two universities in the Campinas' metropolitan area; (2) Ranking the top 100 most accessed websites in each university; (3) Selecting two common websites between both universities.

1. The chosen universities were defined as University 1 (UNIV 1) and University 2 (UNIV 2). The first one offers 20 undergraduate courses and has an average of 2000 students. The second one offers 28 undergraduate courses and has an average of 4000 students.
2. The list with the top 100 most accessed websites in each university did not take into account search engines and social networks due to security policies implemented by each university. The period to collect these data was from October 20th to October 24th, 2014.
3. The parameters for choosing two common websites between both universities were: (a) both websites had to be among the top 20 most accessed websites and (b) their privacy policies should be available in the native language of the students (in this case, Portuguese). From this selection, we got websites from Company 1 (Co 1) and Company 2 (Co 2). The first one is a news website that has been active for over 15 years. It has nationally renowned journalists, friendly interface and great public acceptance.

The second one is known as the market for selling airplane tickets and tourist packages on their online store.

Selection and distribution of students

The chosen students came from different undergraduate courses: engineering, computer science, fashion and social communication (marketing). The selection method was the availability of each group to participate in the experiment. The population of the 198 students took part in this work and the way they are distributed is shown in Table 1. This number is a sample of the students in the campi of the universities, and it represents the different knowledge areas.

Privacy Policy	Co 1		Co 2		Total
	Text Policy	61	Text Policy	31	92
	Label Policy	67	Label Policy	39	106
Gender	UNIV 1		UNIV 2		Total
	Male	90	Male	27	117
	Female	72	Female	9	81
Age	UNIV 1		UNIV 2		Total
	17 to 20 years' old	58	17 to 20 years' old	6	64
	21 to 25 years' old	77	21 to 25 years' old	24	101
	26 to 30 years' old	13	26 to 30 years' old	3	16
	over 30	14	over 30	3	17

Table 1: Distribution of participants.

In our approach was chosen the stratification process dividing members to the population into homogeneous subgroups before sampling. The strata are mutually exclusive where every element into the population was assigned to only one stratum. The strata were selected considering that no population element would be excluded. After this the systematic sampling – using the availability of the students - was applied within each stratum. This approach improves the representativeness of the sample by reducing random sampling error.

Questionnaire

The following questions used in this work were made based on the guidelines for preparation of questionnaires proposed by Lazar et al. (2010). The questions of the questionnaire are:

1. Did you understand the privacy policies?
2. According to the presented privacy policies, were you able to tell if your data were being collected?
3. According to the presented privacy policies, were you able to tell if there was any use of cookies?
4. According to the presented privacy policies, were you able to tell if your geographical localization is collected?
5. According to the presented privacy policies, were you able to tell if your navigation preferences are collected?
6. According to the presented privacy policies, were you able to tell if information regarding your last online purchases are collected?
7. According to the presented privacy policies, were you able to tell if your personal information and documents are collected?
8. According to the presented privacy policies, were you able to tell if your activities on the website are collected?

Every question, except for the first one, had a checkbox similar to the collected data during browsing categories described in the Privacy in the Label Format section. In this sense, each participant should identify and check the boxes related to the data collected during browsing.

Application

The experiment application procedure used some directives presented by Lazar et al. (2010), which are described in the following:

- Make clear that the questionnaire was part of a research on privacy and data security conducted by a research group from Federal University of São Carlos.
- Explain the need of consent of each and every student that was going to be a part of the experiment.
- Make clear which all steps of the experiment were (consent form, handing of privacy policies and the questionnaire) and how long it would take them to finish it (around 20 minutes).
- Hand out the consent form and signatures acknowledgement.
- Inform the students that there would not be a problem if they did not know some terms used in the policies.
- Hand out the privacy policies (textual and labeled) and indicate which website should be visited.
- Hand out the questionnaire.
- Make clear that the questionnaire referred to the students' understanding of the privacy policies.
- Take back all handed out documents (consent form, questionnaire and privacy policy).

Results Analysis and Discussion

The following analysis aims to assess each privacy policy model experimented on this paper. To do so, a template was created with the expected answer – according to the authors' opinions – to each question in the experiment.

The results regarding Company 1 are shown in the two tables. The wrong answers given by the students and their respective percentages in each model are shown in Table 2. The alternative with the most recurrent errors of each privacy policy model is shown in Table 3. In this work, this measure assesses the number of wrong answers, seeking to illustrate what the alternative that showed the highest recurrent error in the students who participated research was.

Questions	Expected Policy			
	Provide the Requested Service		Internal Research and Development	
	Label Privacy (%)	Text Policy (%)	Label Privacy (%)	Text Policy (%)
Question 2	31	9	25	12
Question 3	31	11	29	12
Question 4	33	17	33	12
Question 5	28	9	28	12
Question 6	26	10	23	11
Question 7	64	21	-	-
Question 8	34	12	33	16

Table 2: Statistics for Company 1.

The use of the labeled privacy policy, as shown in Table 2, proved to be a better option to understand the privacy policy of Company 1 than the textual presentation of it. Direct comparison of percentages of each question for both models shows that, in fact, the labeled privacy policy made the understanding of the text easier for the students.

As observed in table 3, the textual model for privacy policies has greatest recurrent error index for all the available questions in the experiment, highlighting the "Market-driven Actions" as the most recurrent error in the textual model, being marked in 5 of the 7 questions that evaluated the intension of the collection in the users' perspective.

Questions	Recurrent Error	
	Label Policy	Text Policy
Question 2	18% Costumer's Profile Assessment	30% Market-driven Actions
Question 3	16% Costumer's Profile Assessment	30% Market-driven Actions
Question 4	11% Costumer's Profile Assessment	24% Market-driven Actions
Question 5	17% Market-driven Actions	31% Market-driven Actions
Question 6	17% Market-driven Actions	31% Market-driven Actions
Question 7	9% Internal Research and Development	21% Provide the Requested Service
	9% Market-driven Actions	21% Costumer's Profile Assessment
Question 8	12% Market-driven Actions	20% Costumer's Profile Assessment

Table 3: Recurrent Error of privacy policies for Company 1.

Questions	Expected Policy	Label Privacy (%)	Text Policy (%)
Question 2	Provide the Requested Service	21	20
	Internal Research and Development	24	17
	Market-driven Actions	24	24
	Telemarketing Actions	18	14
Question 3	Provide the Requested Service	26	25
	Internal Research and Development	22	20
	Costumer's Profile Assessment	26	25
Question 4	No item selected	-	-
Question 5	Internal Research and Development	52	26
Question 6	Internal Research and Development	22	15
	Market-driven Actions	23	29
	Telemarketing Actions	20	17
	Costumer's Profile Assessment	23	23
Question 7	Provide the Requested Service	38	26
	Internal Research and Development	36	21
Question 8	Internal Research and Development	43	36

Table 4: Statistics for Company 2.

Company 2 results are also shown in two tables. The expected answers and their percentages are shown in Table 4, whereas the recurrent error index in each question is shown in Table 5. Table 4 shows that the privacy label format had slightly better results than the textual format for Company 2. Questions 5 and 7, however, clearly shows the superiority of the label approach. The similarity in the answers, according to the authors, is recurring in situations where the textual policy is well written and easy to understand.

Table 5 shows that the textual format had the greatest recurrent error index for Company 2. So it is possible to say that the use of privacy labels gives better results when we consider comprehension and interpretation of Company's 2 privacy policy.

Questions	Recurrent Error	
	Label Policy	Text Policy
Question 2	7% Costumer's Profile Assessment	18% Costumer's Profile Assessment
Question 3	12% Market-driven Actions	17% Market-driven Action
Question 4	19% Provide the Requested Service	22% Provide the Requested Service
	19% Telemarketing Actions	18% Market-driven Action
	19% Costumer's Profile Assessment	24% Costumer's Profile Assessment
Question 5	16% Market-driven Actions	21% Market-driven Actions
Question 6	7% Provide the Requested Service	10% Provide the Requested Service
Question 7	7% Internal Research and Development	21% Costumers Profile Assessment
Question 8	14% Costumer's Profile Assessment	20% Costumer's Profile Assessment

Table 5: Recurrent Error of privacy policies for Company 2.

Conclusions

The main purpose of this paper was to assess the understanding of users regarding privacy policies presented in both textual and labeled models. As a way to measure it, we created a template for each privacy policy. One of the authors that helped create the template is a privacy expert. Each template was created by analyzing each privacy policy and how each company collected data from users. By creating such templates, we had the expected answers and the answers given by the students of the experiment. Therefore, we could measure the convergence and divergence of each student's answer.

The use of labeled privacy policies proved to be the best choice for both companies taken into account in this experiment. The results of Company 1's privacy policy assessment are shown in Tables Table 2 and Table 3. And results of Company 2's privacy policy assessment are shown in Table 4 and Table 5.

The use of natural language for creating privacy policies may induce ambiguity about data collection. This ambiguity comes from verbose texts, which generates disapproval from users regarding such policies, as in McDonald and Cranor (2008). Company 1's privacy policy is a perfect example of what is described in the work of McDonald and Cranor (2008). On the other hand, Company 2's privacy policy makes it easier for the final user to understand which data are being collected, as proposed by Acquisti et al. (2015).

The research developed on this paper also proves that the lack of standard may influence even a more educated public, considering that the research was made in an academic environment, where even with the participant students' knowledge, there was no complete understanding of the terms of the privacy policies. We believe that if the online services adopted a standard for their privacy policies, either textual or visual ones, the results show that the users would have a better understanding, which would make them more aware of which data may be collected, stored and spread.

Finally, we are able to conclude that, for this experiment, the labeled privacy policy model was the best choice to clearly show how the user's data were being collected by each company, as discussed in the Results Analysis and Discussion section.

As future work, we aim to present the results of this paper to companies and collect their opinion regarding the use of the privacy label and its pros and cons.

Acknowledgments

The authors would like to thank the students of the area of privacy and personalization of the master's program of the Federal University of São Carlos, in particular, students Delton Ravagnolli, Leonardo José Lima and Marcelo Castro by significant contributions in the initial phases of this research.

REFERENCES

Acquisti, A., Brandimarte, L., and Loewenstein, G. 2015. "Privacy and human behavior in the age of information," *Science* (347:6221), American Association for the Advancement of Science, pp. 509–514.

Jensen, C., and Potts, C. 2004. "Privacy policies as decision-making tools: an evaluation of online privacy notices," in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems (CHI '04)*, Vienna, Austria: ACM Press, New York, NY, pp. 471–478.

Kelley, P. G., Bresee, J., Cranor, L. F., and Reeder, R. W. 2009. "A nutrition label for privacy," in *Proceedings of the 5th Symposium on Usable Privacy and Security*, p. 4.

Kelley, P. G., Cesca, L., Bresee, J., and Cranor, L. F. 2010. "Standardizing privacy notices: an online study of the nutrition label approach," *Proceedings of the SIGCHI {...}*, pp. 1573–1582 (available at http://dl.acm.org/citation.cfm?id=1753561).

Lazar, J., Feng, J. H., and Hochheiser, H. 2010. *Research methods in human-computer interaction*, Chichester, UK: Wiley Publishing.

McDonald, a M., and Cranor, L. F. 2008. "The Cost of Reading Privacy Policies," *A Journal of Law and Policy for the Information Society* (4:3), pp. 1–22 (available at http://lorrie.cranor.org/pubs/readingPolicyCost-authorDraft.pdf).

McDonald, A. M., Reeder, R. W., Kelley, P. G., and Cranor, L. F. 2009. "A comparative study of online privacy policies and formats," *Proceedings of the 5th Symposium on Usable Privacy and Security - SOUPS '09*, New York, New York, USA: ACM Press, p. 1 (doi: 10.1145/1572532.1572586).